Performance Evaluation and Service Rate Provisioning for a Queue with Fractional Brownian Input ☆

Jiongze Chen[1], Ronald G. Addie[2], Moshe Zukerman[1]

Abstract

The Fractional Brownian motion (fBm) traffic model is important because it captures the self-similar characteristics of Internet traffic, accurately represents traffic generated as an aggregate of many sources, which is a prevalent characteristic of many Internet traffic streams, and, as we show in this paper, it is amenable to analysis. This paper introduces a new, simple, closed-form approximation for the stationary workload distribution (virtual waiting time) of a single server queue fed by an fBm input. Next, an efficient approach for producing a sequence of simulations with finer and finer detail of the fBm process is introduced and applied to demonstrate good agreement between the new formula and the simulation results. This method is necessary in order to ensure that the discrete-time simulation accurately models the continuous-time fBm queueing process. Then we study the limitations of the fBm process as a traffic model using two benchmark models – the Poisson Pareto Burst Process model and a truncated version of the fBm. We determine by numerical experiments the region where the fBm can serve as an accurate traffic model. These experiments show that when the level of multiplexing is sufficient, fBm is an accurate model for the traffic on links in the core of an internet. Using our result for the workload distribution, we derive a closed-form expression for service rate provisioning when the desired blocking probability as a measure of quality of service is given, and apply this result to a range of examples. Finally, we validate our fBm-based overflow probability and link dimensioning formulae using results based on a queue fed by a real traffic trace as a benchmark and demonstrate an advantage for the range of overflow probability below 1% over traffic modelling based on the Markov modulated Poisson process.

Keywords: fractional Brownian motion, long range dependence, Poisson Pareto Burst Process, quality of service (QoS), link capacity dimensioning.

☆The work described in this paper was supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China [CityU 124709].

[1]J. Chen and M. Zukerman are with the Electronic Engineering Department, City University of Hong Kong, Hong Kong SAR (emails: cjz12110@gmail.com, m.zu@cityu.edu.hk)

[2]R. G. Addie is with the Department of Mathematics and Computing, University of Southern Queensland, Australia (email: addie@usq.edu.au)

1. Introduction

It has been well established that Internet traffic is long range dependent (LRD) [1–5]. Furthermore, core and metropolitan Internet links are shared by a large number of users, so, by the central limit theorem, the traffic on such links, which represents multiplexing of traffic generated by many users, can be assumed to follow a Gaussian process for the purpose of performance evaluation and link capacity dimensioning [6]. In addition, Internet traffic is transported based on the store and forward principle, where packets are stored in router buffers before they are forwarded towards their destination. A Gaussian self-similar process, the fractional Brownian motion (fBm), has been widely considered as the model of choice for heavily multiplexed LRD traffic for its simplicity and accuracy. It has been validated, as an accurate model, for such traffic using real measurements as has been reported in numerous publications such as [1, 7–9]. Therefore, a queue fed by fBm input has been considered a fundamentally important model for Internet queueing performance analysis and capacity assignment, and has attracted significant attention [7, 9–19]. However, to date, despite considerable effort, only asymptotic results are available for the queueing performance of fBm queues. Accordingly, it is important to obtain simple and accurate analytical results for fBm queueing performance and service provisioning.

In addition, even though they are only asymptotically accurate, the best results for the performance of a queue fed by fBm available up to now take a form which is not sufficiently explicit to be readily applied for dimensioning. The result presented in this paper is accurate and both mathematically and numerically simple to apply and can be readily used for dimensioning as shown below.

Despite its appeal, the fBm process, as a traffic model, has limitations due to its characteristics and features. As mentioned above, the fBm model is applicable to situations where traffic is highly aggregated. Also, fBm exhibits the feature of *negative traffic* (periods of time during which the total bits arriving is negative), which is not a property of real Internet traffic. For certain ranges of the parameters the impact of this negative traffic is insignificant, and for certain other parameter values it is not. It is important to investigate this feature of the fBm model, to understand how it affects the queueing performance predicted by the fBm model, and to determine whether it causes errors in the application of fBm to modelling the performance of real networks.

The contributions of this paper, their related challenges, and the methods used, are as follows.

(i) A neat and accurate closed-form approximation for the stationary workload distribution of a single server queue fed by an fBm input is provided. The queueing problem of fBm is difficult because it does not lend itself to a classical Markov chain analysis. Therefore new methods have had to be developed for the analysis of an fBm queue. We derive a new formula arising from a new interpretation of an asymptotically accurate formula

for the tail of the distribution in [16], the details of which are shown in Section 2.

(ii) To verify the accuracy of our closed-form approximation, there is a need to simulate an fBm queue. There are two difficulties associated with such simulations. The first difficulty is the adaption of the continuous-time concept of fBm to the discrete-time implementation of a computer simulation. The second one is the large execution time for the generation of long fBm sequences. We overcome these difficulties by a new simulation approach that re-uses a single fBm sequence, as described in Section 3. Then in Section 4, we show the consistency between the simulation and analytical results.

(iii) To show the effect of negative traffic weakness on the queueing performance of an fBm queue, we compare, in Section 5, the queueing performance of the fBm model with that of the Poisson Pareto Burst Process (PPBP) model. Then by means of comparing its queueing performance with that of its truncated counterpart using simulations, we obtain the region where fBm conquers its negative traffic weakness.

(iv) We provide a service provisioning formula in Section 6. To this end, we use Gamma and incomplete Gamma functions to derive the inverse of the performance formula of an fBm queue which leads to a concise service provisioning formula.

(v) We demonstrate the accuracy of our fBm-based closed-form formulae by numerically comparing the overflow probability and link dimensioning results with equivalent results obtained by simulating a single server queue fed by a real traffic trace and fitted Markov modulated Poisson process (MMPP) input traffic in Section 7.

Table 1 lists the definitions of the key parameters used in this paper. A preliminary version of this paper was presented in [20]. The present version is a substantial revision of [20] that includes a new comprehensive discussion on the parameter ranges where the fBm model is applicable, new simulation data to validate the analysis, and various new clarifications, discussions and modifications.

2. A new analytical result for an fBm queue

We consider a single server queue with an infinite buffer fed by an fBm input process, X_t say, with Hurst parameter H, variance in an interval of length 1 equal to σ_1^2 and drift λ. Specifically, the variables σ_1^2 and λ are the variance and the mean of the amount of work arriving during a time interval of length 1. The service rate, denoted by τ [bit/sec.], is assumed constant. Let μ be the mean net input during a time interval of length 1, i.e., $\mu = \lambda - \tau$. For stability we assume $\mu < 0$. Henceforth, a time interval of length 1 will be called 1 second [sec.]. Let Q denote the steady state queue size. Because fBm is a continuous-time process we can not define Q as the limit of a quantity defined by Lindley's equation;

Table 1: Table of notation

Notation	Interpretation
λ	arrival rate of the input traffic
σ_1^2	variance of the input traffic arriving in a time interval of length 1
σ_δ^2	variance of the input traffic during the time interval, δ
σ_t^2	variance of the input traffic during the time interval, t
τ (τ^*)	service rate of the queue (provisioned link capacity)
μ	the mean net input that is equal to $\lambda - \tau$
Q	steady state queue size
H	Hurst parameter
C	normalising constant for an approximation based on the asymptotic formula from [16]
c	normalising constant for the density of the new approximation
γ	tail index of Pareto Dist.
δ	minimum possbible value of Pareto Dist.
$\hat{\lambda}$	traffic intensity of PPBP
ε	overflow probability of the queue

however, making use of the formula of Reich [21], we can define

$$Q = \sup_{t \geq 0} -X_{-t} + \tau t,$$

and since $-X_{-t}+\tau t$ has the same probability measure as $X_t - \tau t$, we can identify Q with the more natural looking expression $\sup_{t\geq 0} X_t - \tau t$.

For the case $H = 0.5$, the complementary distribution function of Q, namely, the probability that the fBm queue size exceeds x, denoted $P(Q > x)$, is well known [22] and is given by

$$P(Q > x) = \exp \frac{2\mu}{\sigma_1^2} x, \quad x \geq 0. \tag{1}$$

For $H \neq 0.5$, despite considerable efforts, there is no exact result for $P(Q > x)$.

Of the known results, a first approximation/bound for the complementary distribution function is from [12]:

$$P(Q > x) \approx \exp\left(-\frac{x^{2-2H}(1-H)^{2H-2}|\mu|^{2H}}{2H^{2H}\sigma_1^2}\right), \quad x \geq 0. \tag{2}$$

The authors of [11] showed that this approximation holds in the sense that

$$\lim_{x\to\infty} \frac{1}{x} \log\left(\text{LHS}/\text{RHS}\right) = 0. \tag{3}$$

A more precise approximation results from [16, Theorem 1, Equation (9)] with $\alpha = 2H$, $\beta = 1$, together with the asymptotic approximation for the tail of a

Normal distribution,

$$1 - \Phi(x) \sim x^{-1} \exp\left(-\frac{x^2}{2}\right),$$

giving

$$P(Q > x) \approx C x^{\frac{2H^2-3H+1}{H}} e^{\left(-\frac{x^{2-2H}(1-H)^{2H-2}|\mu|^{2H}}{2H^{2H}\sigma_1^2}\right)}, \ x \geq 0, \tag{4}$$

where C is a certain constant. This approximation holds in the sense that

$$\lim_{x\to\infty} \text{LHS/RHS} = 1. \tag{5}$$

No method for determining C was provided in [16] although there is an extensive literature connecting C with Pickands' constants, and methods for bounding and approximating this constant (which depends on H) are available [23]. In Section 4 we estimate this constant by numerically fitting (4) to simulation results for the purpose of comparing it with our analytical formula.

While the objective in [16] was to obtain a function that is accurate in the limit of $x \to \infty$, namely satisfies (5), our objective is to obtain an accurate approximation for $P(Q > x)$ for all values of x. If the limiting behaviour of the queueing system as $x \to \infty$ is actually typical even for moderate and small x, and if the Eqs. (4) are a good way to characterise this limiting behaviour, then we can use (4) for our objective as a good approximation for the behaviour of an fBm queue.

However, the particular form of (4) does not lend itself to this strategy because if $0.5 < H < 1$, which is usually the case, the RHS of (4) $\to \infty$ as $x \to 0$, whereas ideally it should tend to the limit 1. Hence, as an approximation, the formula (4) is certain to be inaccurate for small x.

In this paper we revise this approach by supposing that it is not the complementary distribution function, but rather the *density* whose character remains stable near ∞ and is characterised by (4). Note, by Theorem 11.11 in [24], the stationary waiting time distribution of an fBm queue has a density, for all $0 < H < 1$, except for one possible atom, which must occur at the minimal value for which the buffer level has non-zero probability density. Let

$$P(Q > u) \approx \int_u^\infty c f(x) dx. \tag{6}$$

in the sense of (3), where c is a certain constant provided below in (8) and f is a function of the form

$$f(x) = x^\kappa e^{\left(-\frac{x^{2-2H}(1-H)^{2H-2}|\mu|^{2H}}{2H^{2H}\sigma_1^2}\right)}. \tag{7}$$

By taking a derivative of the RHS of (4) we find $\kappa = \frac{1-2H}{H}$. The derivative has another term, but this takes the same form but with a smaller exponent for the initial power of x; omitting this term will not invalidate (3) and hence, since we

seek a function of the form specified in (7) (not a sum of functions each of this form) we must omit the term with the lower exponent of x.

The assumption that the true density takes a form asymptotically like (7) is neither superior, nor inferior, to the assumption that $P(Q > t)$ takes a form like (4). Which of these assumptions is closer to the truth will, in general, vary from case to case. We present evidence below, however, which suggests that (7) provides a more accurate approximation than (4) in the present instance.

The density $cf(x)$ has a finite integral for all values of H; this overcomes the problem identified above in the use of (4) as an approximation. We can find c from

$$\begin{aligned} c^{-1} &= \int_0^\infty f(x)dx \\ &= \int_0^\infty x^{\beta-1}e^{-\alpha x^\nu}dx \\ &= \nu^{-1}\alpha^{-\frac{\beta}{\nu}}\Gamma(\frac{\beta}{\nu}), \end{aligned} \tag{8}$$

where

$$\alpha = \frac{(1-H)^{2H-2}|\mu|^{2H}}{2H^{2H}\sigma_1^2},$$

$\beta = \frac{1-H}{H}$, $\nu = 2 - 2H$ and Γ denotes the Gamma function.

It is feasible that (6) is accurate for small x, and perhaps for all x. The simulation results appear to confirm that this is the case. Another benefit of (6) over (4) is that we have a simple formula for c whereas [16] does not provide a way to derive C.

The new formula is simple to calculate and work with, has no unknown parameters, is asymptotically accurate in the sense of (3) rather than (5) but is potentially accurate over the full range of parameter values, and simulation tests have shown that this is, in fact, the case.

We propose this as a way to characterise the asymptotic behaviour of the buffer level distribution, for an fBm queue, which may have more inherent validity than characterising the behaviour of the model by its complementary distribution function. The fact that this regime appears to hold over a wide range of values, including the range where we normally wish to evaluate the model, is shown below by means of simulations.

3. A sequence of simulations of an fBm queue

There are two challenges facing us in simulating an fBm queue: (1) how to adapt the continuous-time concept of fBm to the discrete-time implementation of a computer simulation that considers the queue size at the endpoints of consecutive intervals each of size Δt [sec.] (2) Generation of long fBm sequences requires a great deal of computer time which becomes prohibitive beyond a certain sequence length. The first challenge forces us to consider a limit, namely, the queueing performance limit as we perform a sequence of simulations where

Δt becomes smaller and smaller. The second challenge can be met if we are able to carry out this sequence of simulations with a single fBm sequence, because this will enormously reduce the computation time required.

There is a significant body of work on how to generate fBm traffic [25–29]. We use the sequence of its incremental process, fractional Brownian noise (fBn), obtained by the Hosking recursive method using the code of Dieker [30]. We re-use this same fBn sequence for each simulation, making use of the self-similarity of the fBm process to enable us to re-interpret the original sequence, after making an appropriate transformation, as a sequence of finer and finer views of the same process. In this way we can be confident that our results are not artefacts of the limited resolution of our simulation.

Note: in addition, a check has also made that the results were not overly dependent on the length of the simulations by repeating the whole process with a longer original fBn sequence.

For each scenario (i.e. a given H value), we use one fBn sequence, $\{U_n\}_{n=0}^{N}$, of length $N = 2^{23}$ samples, which is twice the length of the sequences used in [20]. The sequence for one scenario is independent of the sequences used for the others. The sequence $\{U_n\}_{n=0}^{N}$ is characterized by its Hurst parameter, H, and the variance, v_1, of U_1. By definition, $E(U_1) = 0$. We interpret the time between samples as corresponding to a certain time interval, Δt, which we successively reduce while we re-use the same sequence, to represent the same process at a finer level of detail.

The basic algorithm of our discrete-time queue simulation is *Lindley's equation*,

$$Q_{n+1} = \max(0, Q_n + \widehat{U}_n), \tag{9}$$

in which Q_n denotes the queue length at the end of the nth time-interval and $\widehat{U}_n$ is the amount of work arriving in each interval in the discrete-time approximation to the continuous time model.

We now discuss how *one sequence* of standard fBn can be used to generate a whole family of different discrete-time approximations to the continuous time model, for different choices of sampling interval. For this purpose it is sufficient to set $\widehat{U}_n = s(\Delta t)U_n + m(\Delta t)$ where $s(\Delta t)$ and $m(\Delta t)$ are chosen so that $\widehat{U}_n$ has the appropriate mean and variance. The Hurst parameter of $\{\widehat{U}_n\}$ will then be the same as that of $\{U_n\}$.

According to [1], the variance of work arriving in an interval of length Δt should be $\sigma_1^2(\Delta t)^{2H}$, but the variance of U_1 is v_1, so $s(\Delta t)^2 v_1 = \sigma_1^2(\Delta t)^{2H}$; therefore

$$s(\Delta t) = \frac{(\Delta t)^H \sigma_1}{\sqrt{v_1}}. \tag{10}$$

Similarly, the work arriving in an interval of length Δt should be $\mu\Delta t$ so

$$m(\Delta t) = \mu\Delta t. \tag{11}$$

Thus, using (10) and (11) to derive the family of sequences $\{\widehat{U}_n\}$ from a single fBn sequence $\{U_n\}$, we are able to simulate the fBm queue with a range of different sampling intervals with no need to generate more than one fBn sequence.

This approach is used in the simulations in the next section to ensure that the discretisation of time is not compromising the accuracy of the simulations.

4. Validation of the workload distribution

In this section, we present numerical results that demonstrate the accuracy of our simulation and of our approximation. All simulation results are provided with 95% confidence intervals based on the Student's t-distribution, estimating the standard deviation by dividing the simulation into blocks.

4.1. Validation of the simulation

To validate the simulation, in Figure 1 using the known exact result (1) for the case $H = 0.5$, we demonstrate how the sequence of simulations as described above approach the exact result using only one fBm sequence. We note that in the case $\Delta t = 0.01$, the analytical results are within the simulation confidence intervals.

4.2. Validation of the analytical formula

In Figures 2 – 6, we present results for $H = 0.3$, 0.4, 0.6, 0.7 and 0.8, respectively, that demonstrate the accuracy and robustness of our approximation. In most instances the approximation is within the confidence intervals of the simulation over the entire range and in all instances the discrepancies which do exist are quite small. We also demonstrate that, as expected, the existing asymptotes are not accurate for the full range of parameters in all cases.

4.3. Discussion of results

First of all it should be observed that the formula presented here, based on comparison with simulation, appears to be more accurate than the existing alternative formulae in all the situations which have been tested, up to now.

Second, the new formula has good accuracy for the full range of parameters.

Third, although it might appear that the Hustler-Piterbarg asymptote [16] also provides adequate accuracy over the full practical range of the parameters (for our purposes, for example, its weakness when H and x are small is not practically important), this formula cannot readily be used independently of simulations because it makes use of a constant for which there is no clear method of calculation. In order to use the Hustler-Piterbarg asymptote in the above comparisons, this constant was estimated by simulations, as discussed in Section 2.

5. Constraints on the suitability of the fBm model

Since traffic becomes closer and closer to Gaussian as the degree of aggregation increases [6], and the autocovariance of traffic appears to be approximately the same as fBm [1], for sufficient levels of aggregation, we expect an fBm model

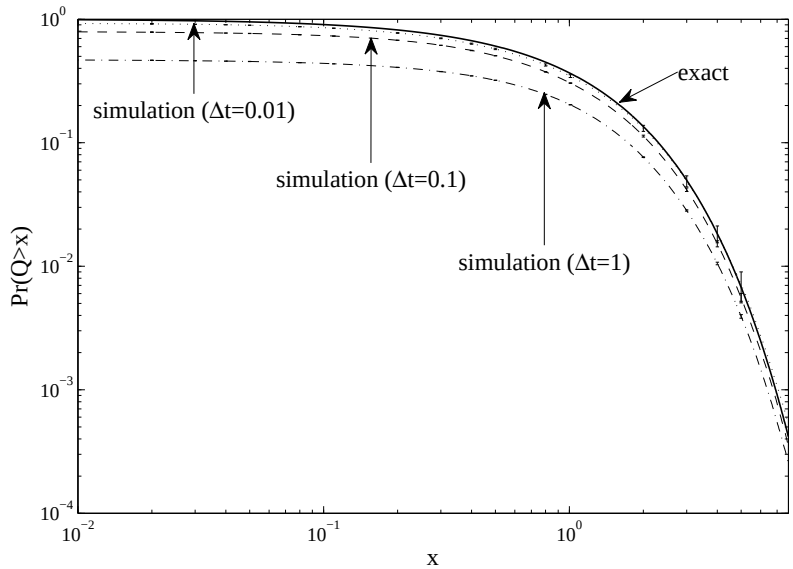

Figure 1: Overflow probabilities based on exact formula vs. simulation results for $H = 0.5, \sigma_1^2 = 1, \mu = -0.5$ with Δt =1, 0.1 and 0.01.

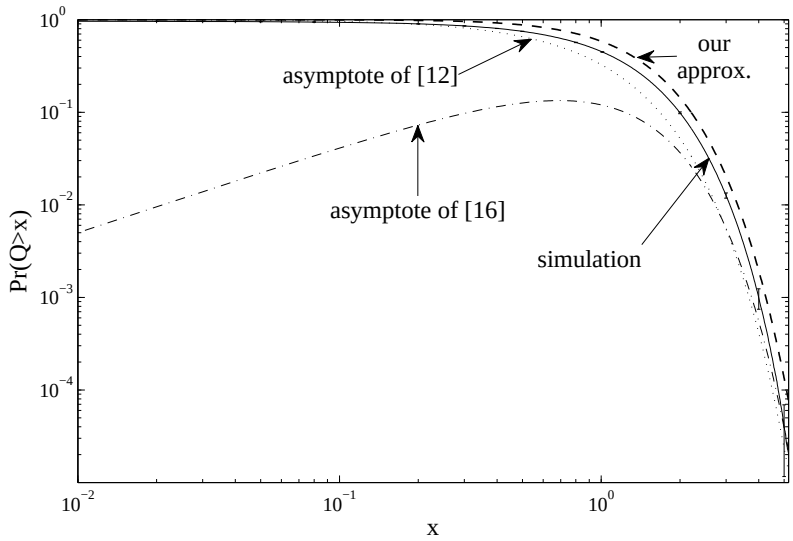

Figure 2: Overflow probabilities based on the asymptotes of [12] and [16], and our approximation vs. simulation results for $H = 0.3, \sigma_1^2 = 1, \mu = -0.5$.

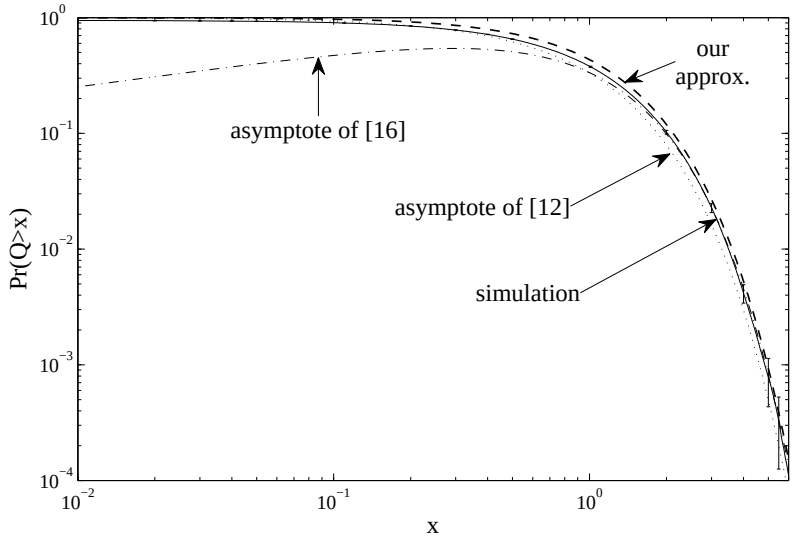

Figure 3: Overflow probabilities based on the asymptotes of [12] and [16], and our approximation vs. simulation results for $H = 0.4, \sigma_1^2 = 1, \mu = -0.5$.

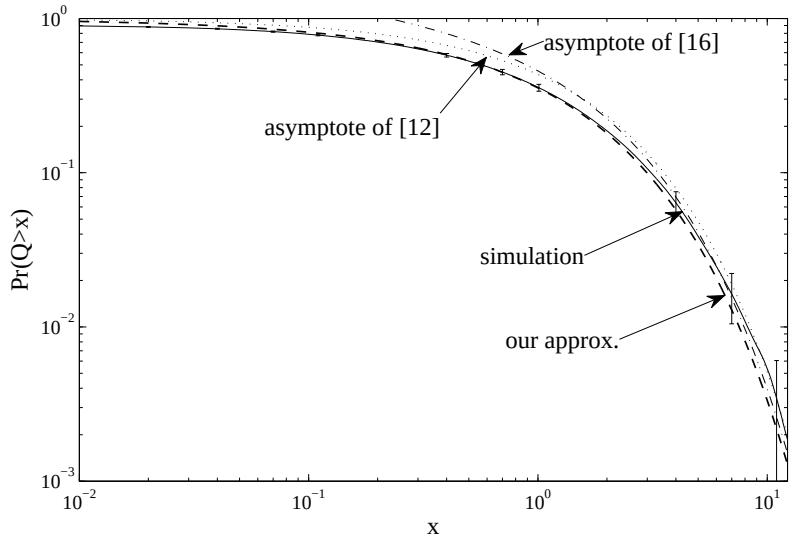

Figure 4: Overflow probabilities based on the asymptotes of [12] and [16], and our approximation vs. simulation results for $H = 0.6, \sigma_1^2 = 1, \mu = -0.5$.

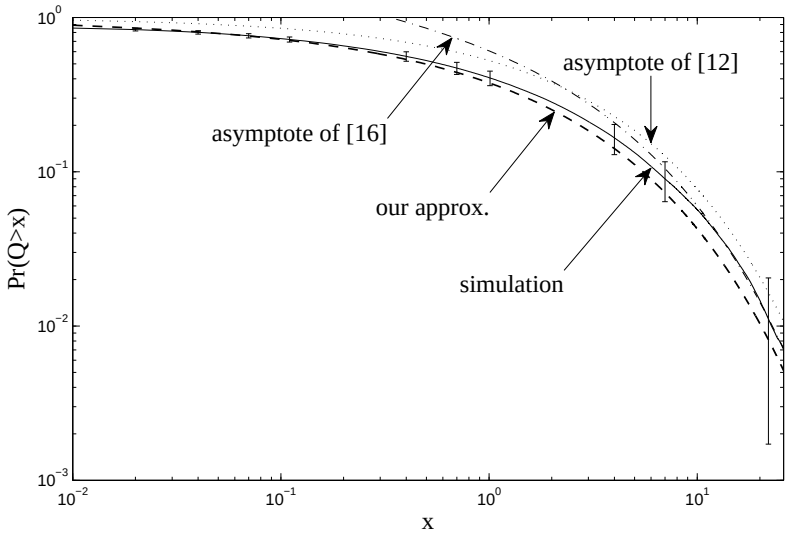

Figure 5: Overflow probabilities based on the asymptotes of [12] and [16], and our approximation vs. simulation results for $H = 0.7, \sigma_1^2 = 1, \mu = -0.5$.

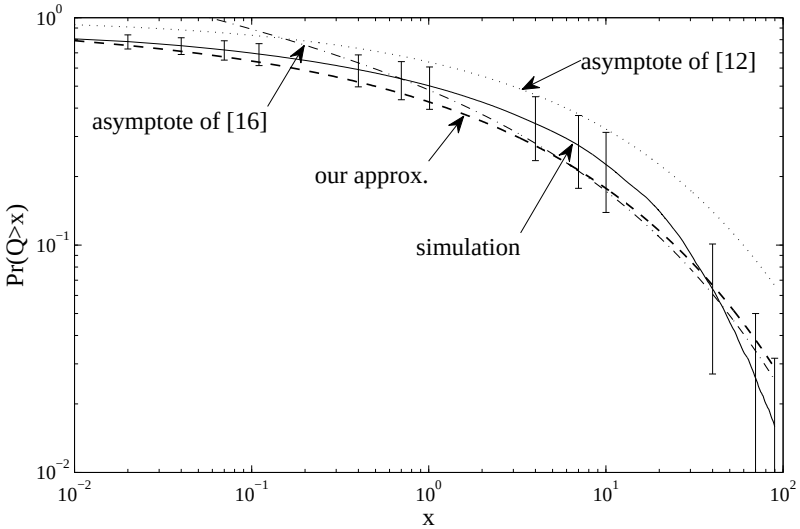

Figure 6: Overflow probabilities based on the asymptotes of [12] and [16], and our approximation vs. simulation results for $H = 0.8, \sigma_1^2 = 1, \mu = -0.5$.

to be suitable for Internet traffic. In this section we carry out numerical comparisons with two important benchmark models. The first is the PPBP model used to answer the question of how high the level of aggregation must be, for the fBm model to be suitable for service rate provisioning and link dimensioning in practice. The second benchmark model used is a truncated version of the fBm that avoids the negative arrivals. It provides insight into the effect of *negative traffic* on the accuracy of the fBm as a traffic model. The fBm model has an unavoidable feature of negative traffic, which is more and more evident as we consider finer and finer detail of the model. This highly counter-intuitive feature of the traffic almost disappears, in practice, if the capacity of the system carrying the traffic is sufficiently large.

5.1. PPBP model

There are several reasons for us to use PPBP for comparison and benchmarking of the fBm model. First, compared with the fBm model, the PPBP, as a traffic model, is more general because it is not limited to cases of heavy multiplexing; so it can help us identify the parameter region where insufficient multiplexing introduces errors when we use fBm as a traffic model. Second, PPBP does not have negative arrivals, therefore it can help us understand the parameter region where the negative arrivals introduce errors when we use fBm to model the traffic. These features of the PPBP make it superior to fBm as a traffic model. Unfortunately, the PPBP does not lend itself to a closed-form analytic solution. In our comparison we use the numerical approach of [31] to generate the queueing performance results for PPBP.

We consider a single server queue fed by a PPBP input. The service rate is denoted by τ [bit/sec.]. The rate at which data is generated during each burst is assumed as a constant, r [bit/sec.]. The arrivals of new bursts follow a Poisson process with rate $\hat{\lambda}$. The time duration of each burst, denoted d, follows a Pareto distribution, the CDF of which is

$$P(d > x) = \begin{cases} \left(\frac{x}{\delta}\right)^{-\gamma}, & x \geq \delta, \\ 1, & \text{otherwise}, \end{cases} \tag{12}$$

where δ is the scale parameter which determines the minimum value of d, and γ is the shape parameter that controls the tail behaviour of the distribution. We have

$$E(d) = \frac{\delta\gamma}{\gamma - 1} \text{ for } \gamma > 1.$$

The mean traffic, λ, is given by:

$$\lambda = \hat{\lambda} r E(d) = \frac{\hat{\lambda} r \delta \gamma}{\gamma - 1} \tag{13}$$

We also have the expression for the variance of the arriving bits during the time

interval, t, in a PPBP from [32]:

$$\sigma_t^2 = \sigma_t^2(\hat{\lambda}, \gamma, \delta, r) = \begin{cases} 2r^2\hat{\lambda}t^2\left(\frac{\delta\gamma}{2(\gamma-1)} - \frac{t}{6}\right), 0 \le t \le \delta, \\ 2r^2\hat{\lambda}\left(\frac{\delta^3\gamma}{6(3-\gamma)} - \frac{\delta^2 t\gamma}{2(2-\gamma)} - \frac{t^{3-\gamma}\delta^\gamma}{2(1-\gamma)(2-\gamma)(3-\gamma)}\right), \quad t > \delta. \end{cases} \tag{14}$$

5.2. Comparison between fBm and PPBP

We compare the fBm and PPBP models using the overflow probabilities obtained by our approximation as described above and the Quasi-Stationary (QS) algorithm for the PPBP queue of [31]. To achieve a fair comparison, we aim to match the key statistical characteristics of the two processes. In the following we describe how we choose an fBm process, described by the three parameters 1) the mean net input, μ, 2) the Hurst parameter, H, and 3) the variance, σ_1^2, for a given PPBP process described by $\hat{\lambda}$, r, δ and γ.

First, by Equation (13) we obtain

$$\mu = \frac{\hat{\lambda}r\delta\gamma}{\gamma - 1} - \tau.$$

Second, the value of the Hurst parameter H is obtained by the formula [32]

$$H = \frac{3-\gamma}{2}.$$

Finally, we derive an equation for the variance σ_1^2 of the fBm process in terms of the parameters of the PPBP process such that the variance-time curves of the two processes are close to each other. To achieve a close match of the variance-time curves, we first note that the variance-time curve σ_t^2 of fBm is given by

$$\sigma_t^2 = \sigma_1^2 t^{2H} \tag{15}$$

Next, we choose an appropriate t_1 value such that

$$\sigma_{t_1}^2(\text{fBm}) = \sigma_{t_1}^2(\text{PPBP})$$

where the right-hand side of this equation is obtained by (14). If we have made a good choice of t_1, then we can obtain the required fBm parameter, i.e. the variance σ_1^2, in terms of H by (15). Then, the fBm variance-time curve can be obtained using (15) as

$$\sigma_t^2 = \sigma_{t_1}^2\left(\frac{t}{t_1}\right)^{2H}. \tag{16}$$

The remaining problem is therefore the choice of t_1. This is done by observing the behaviour of the variance-time curves for various choices of the t_1 parameters. Let us illustrate our approach using the following example. We set the PPBP parameters as follows: $\gamma = 1.5$, $\delta = 1$ and $r^2\hat{\lambda} = 1$. Note that by (14), the family of PPBP processes that obey the relationship $r^2\hat{\lambda} = 1$, all have the

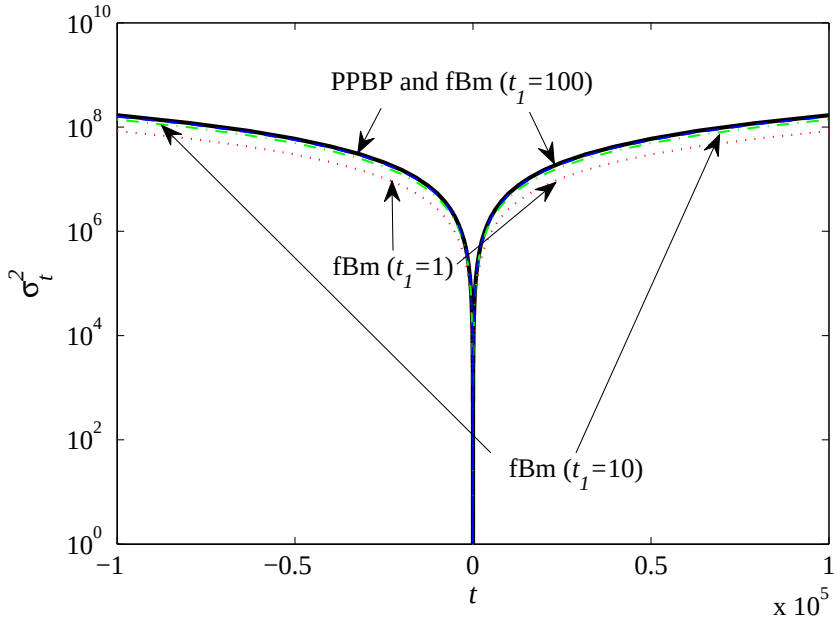

Figure 7: Variance-time curve of fBm vs. PPBP for $\gamma = 1.5$ and $\delta = 1$

same variance-time curve. These are sufficient to uniquely define the variance-time curve of the PPBP. Then we choose three values for t_1: 1, 10, and 100. To obtain the corresponding fBm time curves, we use $H = (3-\gamma)/2 = 0.75$. Then having H and t_1 we use (16) to uniquely determine the variance-time curve. The three corresponding variance-time curves of the fBm are plotted in Figure 7. We observe that for reasonably large time scales choosing $t_1 \geq 10$ will give reasonably accurate match, so we choose $t_1 = 10$.

We set $\gamma = 1.5$ so that $H = 0.75$, which is a realistic value for Internet traffic. The values for mean net input, μ, and r are the same here as in [31], with

$$\mu = -\frac{3\sigma_\delta(\hat{\lambda}, \gamma, \delta, r)}{\delta} \text{ and } r = 1.$$

The values of $\hat{\lambda}$ are chosen as 1, 1024 and 128000 to model steady growth of traffic intensity, as might be expected in a real network. Since it is known *theoretically* that the Poisson-Pareto burst model converges to the behaviour of a Gaussian model with almost the same variance-time curve as fractional Brownian motion as $\lambda \to \infty$ [33], checking that this convergence holds using our new formula for the stationary buffer level in an fBm queue provides further confirmation of the validity of the analyses of both models. To undertake this test thoroughly we should check that it holds over the full range of choices for the parameters of these models. It is not necessary to repeat the comparison for

choices of parameters which are only trivially different (in the sense of scaling), and consequently the three figures 8–10 are sufficient to explore a fairly full range of parameter values.

These figures show a comparison of overflow probabilities of PPBP model and fBm model for $\delta = 0.1$, 0.2 and 0.5, for figures 8, 9, 10, respectively. We choose comparatively small δ because it represents the minimum burst duration which is small in practice because bursts represent application flows which can be relatively small. In addition, note that the smaller the δ is, the less the discrepancy between the variance-time curves of the PPBP model and fBm. The three parameters of the PPBP model, δ, $\hat{\lambda}$ and r, linearly depend on the time unit selected. Increasing these three parameters at the same rate is equivalent merely to a change of time unit, and hence their queueing behaviours are identical.

The x-axis of the plots in Figure 8– Figure 10 is in units of x/σ_1 where x is the buffer threshold level. This is because when the central limit theorem is applied to these PPBP processes convergence to a certain common Gaussian process will occur only if the arriving work is measured in these (or some equivalent) units.

As expected, we see the convergence between the two models for high aggregation ($\hat{\lambda} = 128000$) on a large buffer threshold, for all three graphs. Compared with the fBm model, the PPBP model has the following differences: 1) it is a non-Gaussian process, 2) it has a slightly difference variance-time curve (with the difference reducing as $\delta \to 0$) and 3) it contains no negative arrivals. According to the central limit theorem, PPBP tends to be Gaussian when $\hat{\lambda}$ is very large. Also the difference of the variance-time curves of fBm and PPBP is minimised by choosing a suitable value of t_1—which has been explained in detail in the previous part of this section. So, we consider negative arrivals feature of the fBm model as the main reason for the disagreements in the range associated with small buffer thresholds even for large $\hat{\lambda}$, i.e. $\hat{\lambda} = 128000$. Intuitively, the less buffer there is in the system, the higher the effect of negative arrivals will be. In the following we focus our attention on the effect of the negative arrivals feature of fBm.

5.3. The Problem of Negative Traffic

Unavoidably, negative arrivals phenomenon is part of the fBm model. It is not desirable because it is inconsistent with traffic behaviour in the real world. In this section, we study the effect of negative arrivals by comparing the performance of a queue fed by an fBm input with its truncated counterpart that does not exhibit negative arrivals. In particular, we now use as benchmark a single server queue whose input is a truncated version of the fBm model so that negative arrivals are avoided. A particularly important issue that we address in this section is identifying the range of parameters for which the adverse effect of the negative arrivals on the overflow probability is negligible which makes the fBm model useful in practice.

We find that the influence of negative arrivals is highly related to the ratio of λ and σ_1, which is intuitively understandable. A series of experiments based on

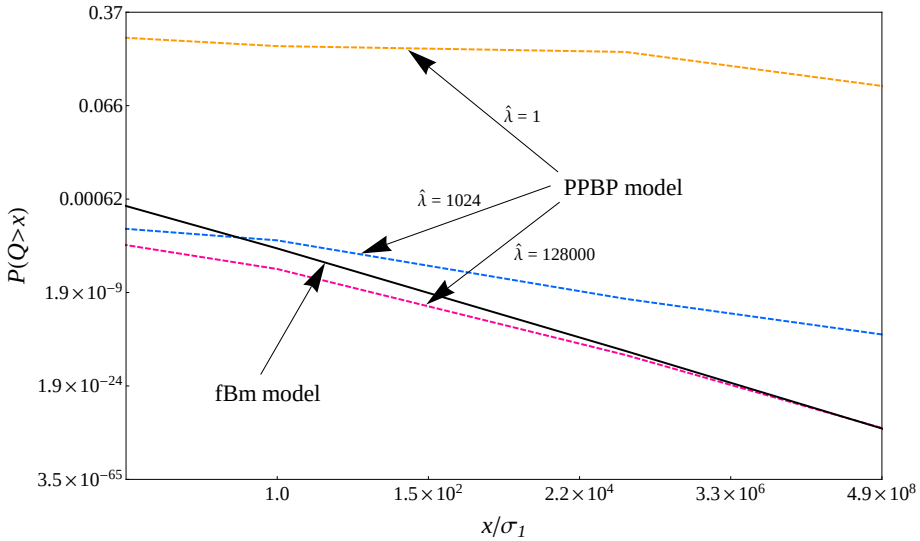

Figure 8: Overflow probabilities for PPBP model with various arrival rates $\hat{\lambda}$, estimated by the QS algorithm vs. the fBm model by our approximation $\delta = 0.1$

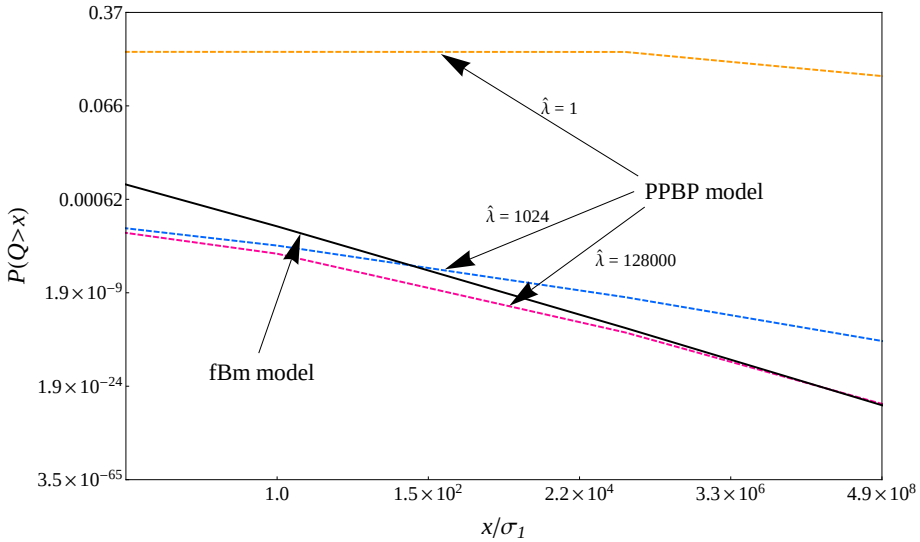

Figure 9: Overflow probabilities for the PPBP model with various arrival rates $\hat{\lambda}$, estimated by the QS algorithm vs. the fBm model by our approximation for $\delta = 0.2$

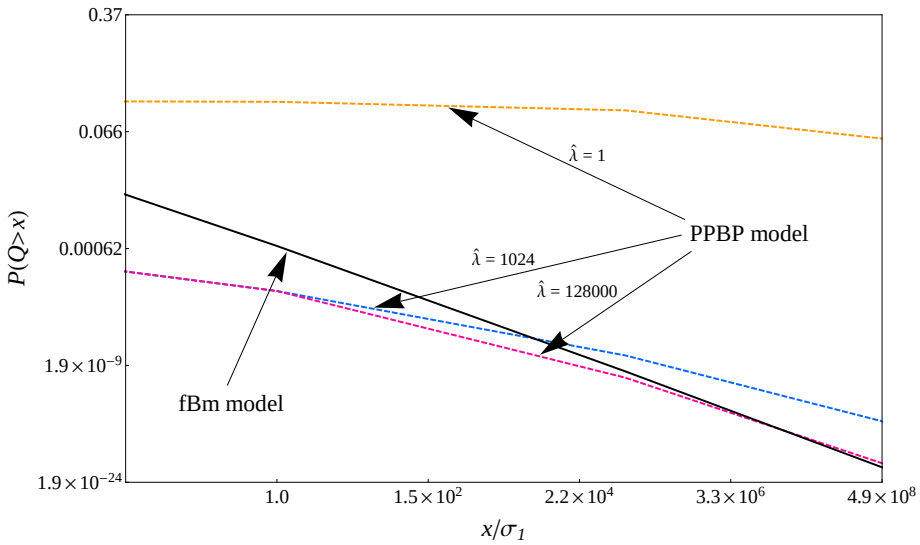

Figure 10: Overflow probabilities for the PPBP model with various arrival rates $\hat{\lambda}$, estimated by the QS algorithm vs. the fBm model by our approximation $\delta = 0.5$

simulations has been done in order to obtain a bound by which we are able to know if the fBm is a realistic traffic model. Different H values — i.e. $H = 0.3$, 0.4, 0.5, 0.6, 0.7 and 0.8 — are chosen for each experiment, where we use various ratios of λ and σ_1 (with increment as 0.1) for both queues fed by an fBm input and its truncated counterpart. When the difference of the overflow probabilities of the two inputs is below a certain level (here we choose 10%), then we say the fBm model is realistic with the certain ratio; otherwise, we say it is unrealistic. By recording all the results for $H = 0.3 - 0.8$, we obtain Figure 11, where the regions of when fBm is a realistic model or not have been shown. We see that as the arrival rate increases, the effect of negative arrivals diminishes. For convenience, a curve involving λ/σ_1 and H has been drawn as a threshold.

In addition, notice that $P(Q > 0)$ equals 1 regardless of whether the fBm input is truncated or not. This is not valid for real traffic, which may exhibit empty buffers for an extended period of time and hence reveals a weakness of the fBm model – it is not a suitable traffic model for cases where the buffer size is small. This weakness means that fBm is usually not applicable to an all-optical network, where buffering light is difficult; however the fBm model is relevant to the current Internet, where the buffer sizes of routers are very large.

5.4. Discussion

The fBm model is not universally appropriate to Internet traffic. An important weakness of this model is that for certain ranges of the parameters it exhibits the phenomenon of *negative traffic* to an excess. A small amount of negative traffic will always be present in the fBm model, but since real networks

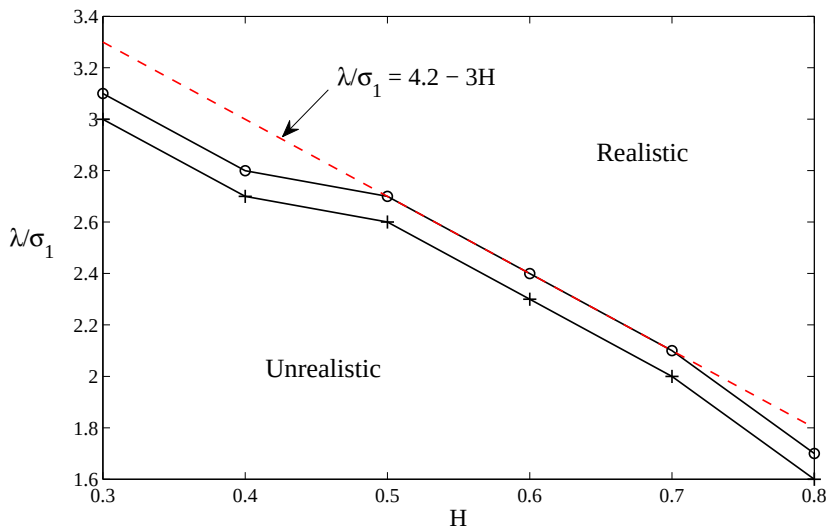

Figure 11: Regions where fBm is a realistic model or unrealistic model due to the presence of negative traffic.

have no negative traffic at all, it is inappropriate to use a model which has large amounts of it. When σ_1 is large relative to λ, or when $H \ll 0.5$, and especially when both of these conditions hold, the model will exhibit large amounts of negative traffic when observed at small time scales.

It may be possible to re-interpret an fBm model in a manner which alleviates this problem and enables us to use this model for a wider range of parameters. Since the dimensioning formula, which will be introduced in the next section, has such a natural interpretation, it may be useful to develop methods whereby it can be applied to a wider range of situations. However, in the absence of such a strategy it will be necessary to confine the use of the fBm model to the region in Figure 11 labelled as *Realistic*. This constraint on σ_1/λ means that the fBm model is applicable only to core links of networks, where the traffic is relatively smooth.

6. Service rate provisioning

Having validated the analytical formula we have derived for the workload distribution of a queue fed by fBm traffic, we are now in a position to evaluate the capacity required to serve a queue fed by fBm traffic such that the overflow probability for a given queue threshold is below a given margin. In the following theorem, we provide a service rate provisioning formula with which the required service rate can be calculated to meet a certain QoS measures-overflow probability and queue threshold.

Proposition 1. *Consider a single server queue fed by an fBm input characterised by the mean rate λ, the variance σ_1^2, and the Hurst parameter H. Let ε be the overflow probability of an fBm queue. Using the notations of (6), we assume that*

$$\varepsilon = P(Q > q) = \int_q^\infty cf(x)dx. \tag{17}$$

That is, the approximation (6) is assumed to be exact. Then, the capacity which exactly meets the performance requirement is

$$\tau^* = \lambda + \left(\frac{2H^{2H} G^{-1}\left((2H)^{-1}, \varepsilon\right)}{(1-H)^{2H-2}} \right)^{\frac{1}{2H}} \sigma_1^{\frac{1}{H}} q^{-\frac{1-H}{H}}, \tag{18}$$

in which $G^{-1}(\alpha, y)$ is the inverse regularised incomplete Gamma function defined by the property

$$G^{-1}(\alpha, G(\alpha, x)) = G(\alpha, G^{-1}(\alpha, x)) = x, \tag{19}$$

where $G(a, x)$ is the regularised incomplete Gamma function

$$G(a, x) = \frac{\int_x^\infty t^{a-1} e^{-t} dt}{\Gamma(a)}. \tag{20}$$

Proof. By substituting kt for t of (20), we find

$$\Gamma(a) G(a, x) = k^a \int_{x/k}^\infty t^{a-1} e^{-kt} dt.$$

Then we can derive a dimensioning formula from (17) as follows:

$$\begin{aligned} \varepsilon &= \int_q^\infty cf(x)dx \\ &= \frac{\nu \alpha^{\frac{\beta}{\nu}}}{\Gamma\left(\frac{\beta}{\nu}\right)} \int_q^\infty y^{\beta-1} e^{-\alpha y^\nu} dy \end{aligned}$$

which, using the substitution $u = y^\nu$, in which case $du = \nu y^{\nu-1} dy$, or putting it another way, $du = \nu u^{\frac{\nu-1}{\nu}} dy$

$$\begin{aligned} &= \frac{\alpha^{\frac{\beta}{\nu}}}{\Gamma\left(\frac{\beta}{\nu}\right)} \int_{q^\nu}^\infty u^{\frac{\beta-1}{\nu}} u^{\frac{1-\nu}{\nu}} e^{-\alpha u} du \\ &= \frac{\alpha^{\frac{\beta}{\nu}}}{\Gamma\left(\frac{\beta}{\nu}\right)} \int_{q^\nu}^\infty u^{\frac{\beta}{\nu}-1} e^{-\alpha u} du \\ &= G\left(\frac{1}{2H}, \alpha q^\nu\right). \end{aligned} \tag{21}$$

Recall that $\nu = 2 - 2H$, with (19) we obtain

$$\alpha = G^{-1}\left(\frac{1}{2H}, \varepsilon\right) q^{2H-2}. \tag{22}$$

Let us now use the original defining equation for α, which expresses it in terms of μ, to determine a formula for μ and hence then τ^*:

$$\begin{aligned} \alpha &= \frac{(1-H)^{2H-2}|\mu|^{2H}}{2H^{2H}\sigma_1^2} \\ \Longrightarrow (1-H)^{2H-2}|\mu|^{2H} &= 2\alpha H^{2H}\sigma_1^2 \\ \Longrightarrow |\mu| &= \left(\frac{2\alpha H^{2H}\sigma_1^2}{(1-H)^{2H-2}}\right)^{\frac{1}{2H}}. \end{aligned}$$

Now since $\mu = \lambda - \tau^*$, where τ^* is the capacity and λ denotes the mean rate of the input traffic,

$$\tau^* - \lambda = \left(\frac{2\alpha H^{2H}\sigma_1^2}{(1-H)^{2H-2}}\right)^{\frac{1}{2H}}.$$

By elementary algebra operation, (18) is proved. □

Proposition 1 provides a simple and elegant formula for service rate provisioning for a queue loaded by fBm traffic on the basis of the assumption that (6) holds. The extensive simulations used to validate the approximation (6) therefore validate also our service rate provisioning formula. From (18) simple relationships can be observed. One observation is that the required spare capacity beyond the arrival rate (λ) is independent of λ. Another simple observation is that the spare capacity required is proportional to $\sigma_1^{2/2H}$. In the following, we provide numerical results illustrating the application of this formula.

6.1. Numerical results

Numerical results of service rate provisioning for a range of examples using the method developed above are presented in Figures 12–15. The results presented are for the total capacity τ^*, but it is important to keep in mind the concept of spare capacity $\tau^* - \lambda$. In all the figures we set $\lambda = 1$. This could represent a data rate in order of Gb/s, e.g., one OC-192 or OC-768 with rate of 10 Gb/s or 40 Gb/s. Then $q = 0.1$ would represent a QoS measure of 100 ms.

In Figure 12, we illustrate the total capacity τ^* required as a function of the second QoS measure ε within the range 10^{-5} – 1. As expected the spare capacity required, i.e. $\tau^* - 1$, reduces with the QoS measure ε. For the case $\lambda = 1$, the spare capacity $\tau^* - 1$ approaches zero as the QoS is further and further relaxed.

In Figure 13, we illustrate that the total capacity τ^* increases with the Hurst parameter H. This is also expected as stronger correlation in the traffic stream means higher queueing delay. Notice, however, that the increase remains bounded as $H \to 1$.

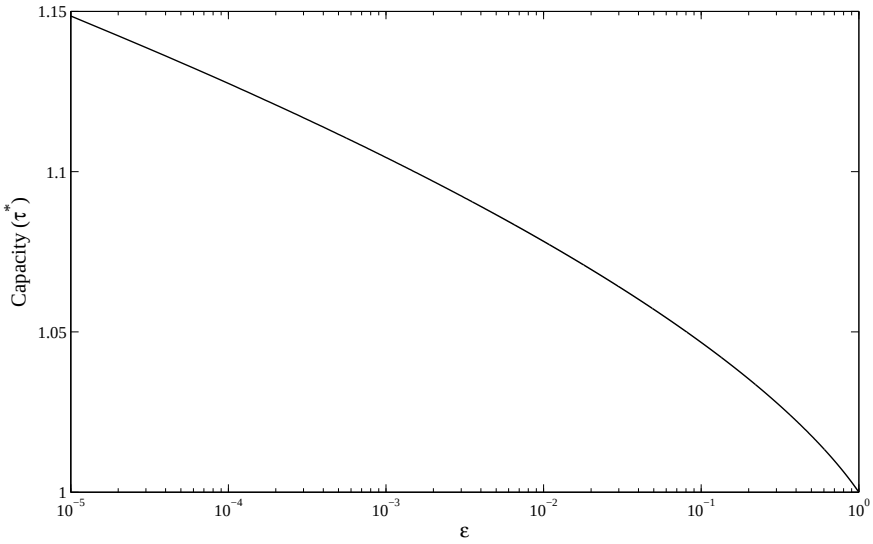

Figure 12: Capacity vs. ε by (18) for $H = 0.8$, $\sigma_1 = 0.05$, $\lambda = 1$ and $q = 0.1$.

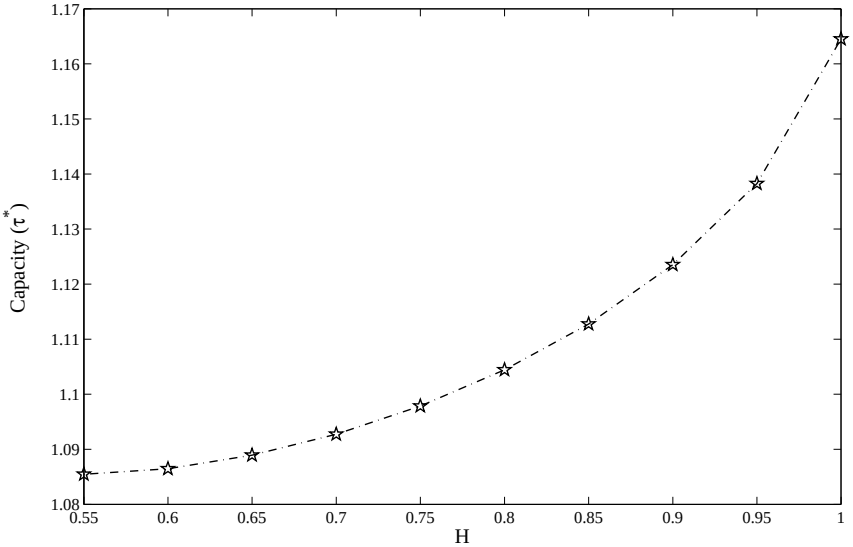

Figure 13: Capacity vs. H by (18) for $\sigma_1 = 0.05$, $\lambda = 1$, $q = 0.1$ and $\varepsilon = 0.001$.

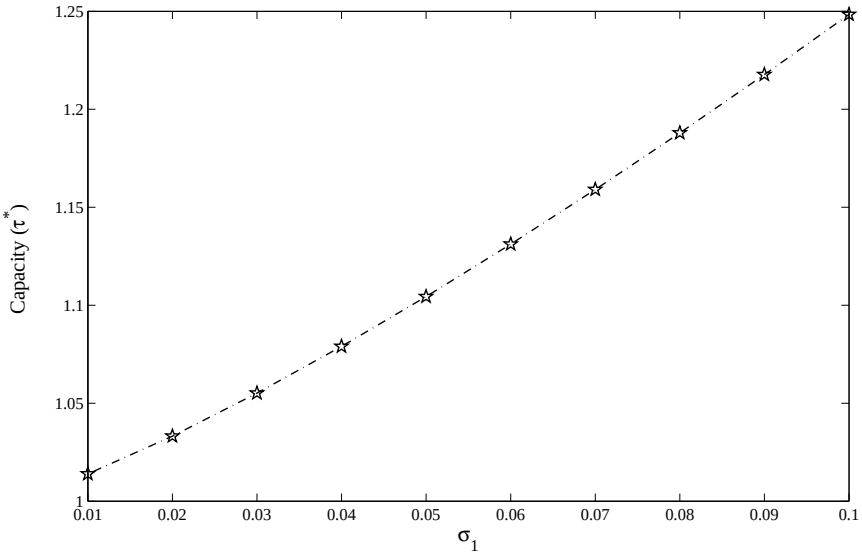

Figure 14: Capacity vs. σ_1 by (18) for $H = 0.8$, $\lambda = 1$, $q = 0.1$ and $\varepsilon = 0.001$.

In Figure 14, we illustrate that the total capacity τ^* increases with the standard deviation parameter σ_1. This is also expected as stronger variation in the traffic stream, normally, means higher queueing delay.

In Figure 15, we illustrate that the total capacity τ^* decreases as the QoS parameter q increases. Thus, allowing more delay and more buffering relaxes delay requirement, so less capacity is required. The required spare capacity approaches zero as the queue size approaches infinity. This is consistent with known results in elementary Markovian queues.

7. Validation based on real traffic data

In this section, we provide evidence based on real traffic data for the accuracy of our fBm-based overflow probability approximation and the dimensioning formula. We also describe a useful process for fitting the parameters which are then used in our closed-form formulae. We also provide a comparison with a short range dependent (SRD) model, in particular, the Markov modulated Poisson process (MMPP) [34, 35], and illustrate the benefit of using the fBm.

7.1. Dataset description

A real traffic input data stream is selected from an IP traffic trace provided by CAIDA's equinix-sanjose monitors on high-speed Internet backbone links. The details provided in [36] enable the reader to reproduce the complete traffic trace that we use here. It contains around 1.478×10^9 packets with total size of about 1 Terabyte that were captured within one hour on February 16 2012.

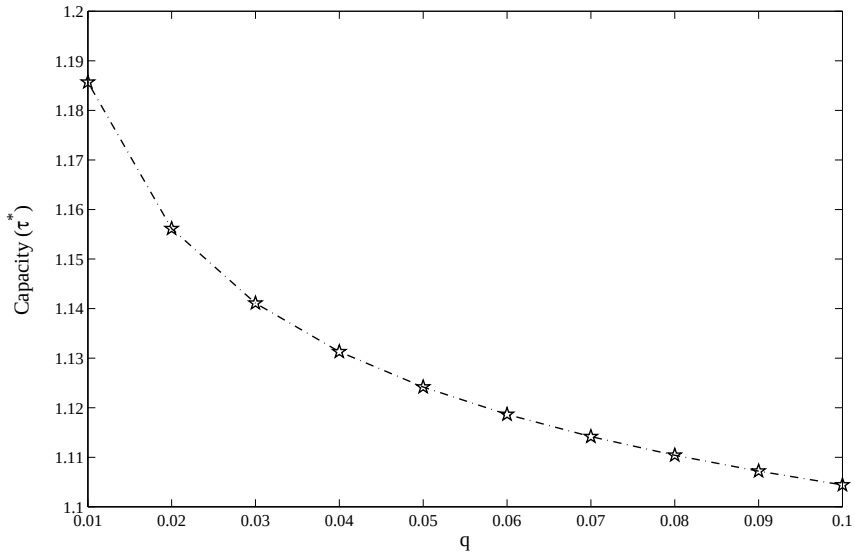

Figure 15: Capacity vs. q by (18) for $H = 0.8$, $\sigma_1 = 0.05$, $\lambda = 1$ and $\varepsilon = 0.001$.

Let the one hour, during which the traffic data is captured, be divided into consecutive fixed time intervals, each of duration t [sec.]. Let S_n be the total traffic [Mbytes] that arrives within the time interval $[(n-1)t, nt)$.

7.2. Fitting fBm parameters to the real measurements

In order to evaluate queueing performance or perform link dimensioning for an fBm queue, we first need to set the values of its parameters, namely, the mean rate (λ), the Hurst parameter (H) and the variance of the arriving workload in an interval of length 1 (σ_1^2). In the following, we show how to fit these three parameters to a real traffic input data stream.

To begin with, the mean rate, λ, can be estimated as the mean of S_n for $t = 1$ [sec.]. For the present case, λ is set to approx 270.87 [Mbyte/sec.]. The values of H and σ_1^2 are estimated by fitting the variance-time curve of S_n. Figure 16 illustrates how we obtain H and σ_1. We first find the variances of S_n for various t values, and then fit a variance-time curve to them in a log-log plot. For this case, we choose the following t values: $t = 0.001, 0.01, \ldots, 100$ [sec.]. A straight fitted line gives a slope as 1.723, which is larger than 1, and therefore provides an evidence of the LRD phenomenon for the real traffic. Since the slope of the v-t curve is equal to $2H$, the value of H can be obtained as 0.8615. More sophisticated methods for fitting H are available, such as the Whittle estimator [37] and the Abry-Veitch wavelet method [38, 39]; however in the present context it is more useful to check that the variance-time curve takes the expected form (linear in a log-log plot). Meanwhile, the estimation of σ_1^2 is found from the v-t curve for $t = 1$, that is $\sigma_1^2 = 457.09$. Therefore, the three fBm parameters are

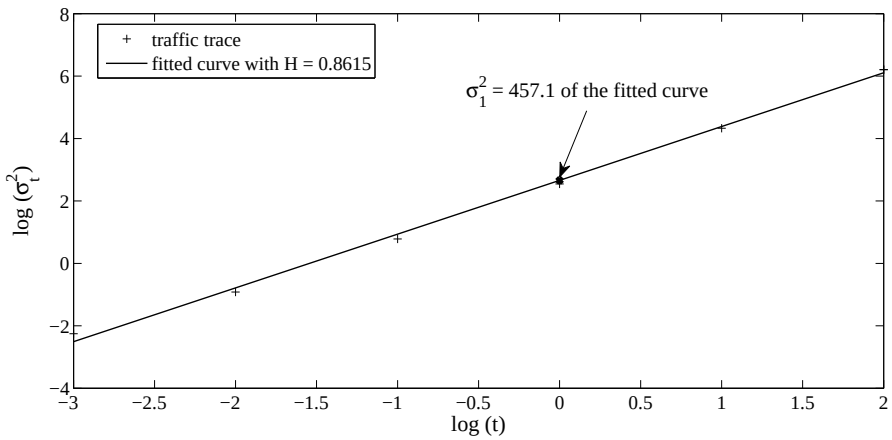

Figure 16: Fitting the variance-time curve.

set as follows: $\lambda = 270.87$, $H = 0.8615$ and $\sigma_1^2 = 457.09$. As the three fBm parameters are available, we can use the closed-form equations (6) and (18) to obtain the overflow probability and to perform link dimensioning, respectively.

7.3. Comparisons between fBm, MMPP and real measurements

We now consider three independent single server queues. They all have the same service rate of τ^*, which is initially set at $\tau^* = 320$ [MByte/sec.]. The input traffic into the first queue is the real data input traffic stream described in Section 7.1. The input traffic into the second queue is the fitted fBm process as described in Section 7.2. The input traffic into the third queue is based on the MMPP process obtained by Algorithm LAMBDA [35]. In the following, we provide information on how we generate the MMPP input traffic so that its parameters are fitted to the real traffic input data stream.

The MMPP input traffic is formed by the MMPP arrival process with general packet size distribution. Algorithm LAMBDA is applied to determine the MMPP state and state transition matrix, by which the MMPP process is generated. Detailed steps for fitting real traffic data can be found in Section II-B of [35]. Also a matlab code for Algorithm LAMBDA is available in [40], by which we obtain 12 states and the corresponding state transition matrix from the real traffic data. For the packet size distribution, we use the actual packet size distribution as the real traffic trace. Since no analytical result is available for a single server queue with MMPP input and general nonparametric service time distribution, the results of queueing performance for MMPP are obtained by simulation.

In the following, we compare the overflow probabilities and link dimensioning results for the above-mentioned three queues. The results for the fBm queue are obtained based on our approximation of the overflow probabilities (6) and the dimensioning formula (18). The results for the two equivalent queues fed

by the fitted MMPP and the actual trace are obtained by simulations. Because the arrival process of both cases is not Poisson, a Poisson inspector is used to measure the queue size during the simulation.

Figure 17 presents the overflow probabilities for the three queues. As mentioned, $\tau*$ is set as 320 [MByte/sec.], which is close to $\lambda + 3\sigma_1$. During the simulations for the queues fed by the real traffic trace and the MMPP input traffic, we independently computed overflow probabilities for each 10 minutes' period and then used them to obtain 90% confidence intervals based on Student's t-distribution, which are presented as the coloured area in the figure. The results from our fBm formula are within the confidence interval of the simulation results of real traces for $q > 0.1$ [MBytes]. The explanation of the differences between fBm formula and real trace simulation results for small q is the existence of negative traffic, which has more effect on small queue size. On the other hand, the simulation results of the queues fed by the real trace and the MMPP input traffic are close to each other for $q < 0.1$ [MBytes]; however, there is almost no overlap between their confidence intervals for $q > 0.1$ [MBytes]. This is explained by the fact that MMPP is SRD while fBm is LRD.

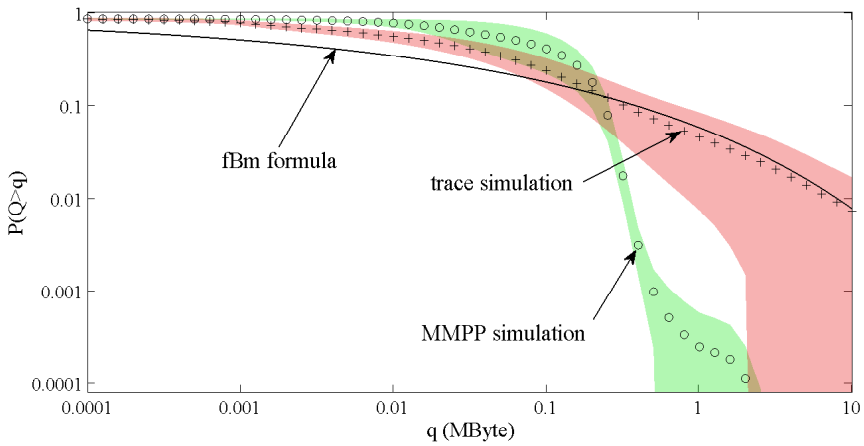

Figure 17: Comparison of overflow probabilities for fBm and MMPP models, and real measurements for a link with capacity $\tau^* = 320$ [MByte/sec.].

Next we consider the link dimensioning problem. For each one of the queue, we aim to find the least link capacity τ^* that ensures that the overflow probability will not exceed 0.001. For the fBm queue, this is achieved based on the dimensioning formula 18. For the other two queues, based on the MMPP input and the input from the real trace, this is achieved by bisection on $\tau*$ using repeated simulations. As shown in Figure 18, the fBm model leads to reasonably accurate dimensioning, with error of no more than 5% and appears to outperform MMPP for the full range of the queue threshold. Noticing that in Figure 17 we observe better accuracy of MMPP for small q values, the explanation of the results presented in Figure 18 is that in Figure 18 we are aiming at much smaller overflow probabilities than the ones for which MMPP performs well in

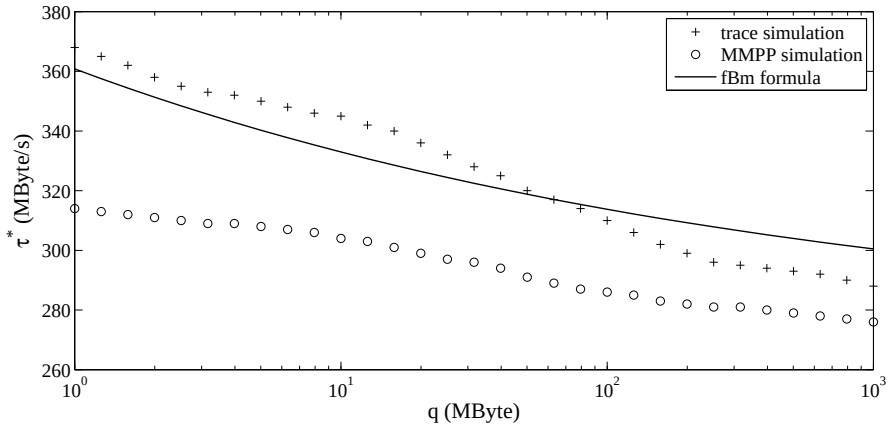

Figure 18: Comparison of the least link capacities for fBm and MMPP models, and real measurements for the overflow probability $\varepsilon \leq 0.001$.

Figure 17. We have noticed from the results in Figure 17 that MMPP significantly underestimates the overflow probability in the range where the overflow probability is below 1%; therefore it also underestimates the required capacity as illustrated in Figure 18. In sum, we demonstrate the accuracy and usefulness of our easy-to-use fBm-based overflow probability approximation and the dimensioning formulae.

8. Conclusion

We have considered a queue fed by an fBm input and derived new results for queueing performance and service rate provisioning. We have also described an efficient approach to simulate such a queue. Agreement between the analytical and the simulation results has been demonstrated and a numerical comparison with existing asymptotes has been presented. Constraints on applying the fBm model so as to use it for link dimensioning have been studied, and we have observed that our formulae can be applied in practical scenarios to network link dimensioning where traffic is heavily multiplexed, and the buffer size and the system service rate are sufficiently large. We have also presented numerical results for a range of examples for link dimensioning based on our queueing analysis. We have demonstrated the accuracy and usefulness of our fBm-based overflow probability and link dimensioning formulae by comparisons with simulation results of a single server queue fed by a real traffic trace, and have shown benefit over MMPP modelling when the overflow probability is below 1%.

Acknowledgment

The authors would like to thank Ilkka Norros and Ward Whitt for advice on the relevant literature.

References

[1] W. E. Leland, M. S. Taqqu, W. Willinger, D. V. Wilson, On the self-similar nature of ethernet traffic (extended version), IEEE/ACM Trans. Networking 2 (1) (1994) 1–15.

[2] R. G. Addie, M. Zukerman, T. Neame, Fractal traffic: measurements, modelling and performance evaluation, in: Proc. IEEE INFOCOM '95, Vol. 3, 1995, pp. 977–984.

[3] J. Beran, R. Sherman, M. S. Taqqu, W. Willinger, Long-range dependence in variable-bit-rate video traffic, IEEE Trans. Commun. 43 (2, 3, 4) (1995) 1566–1579.

[4] V. Paxson, S. Floyd, Wide-area traffic: the failure of Poisson modelling, IEEE/ACM Trans. Networking 3 (3) (1995) 226–244.

[5] M. E. Crovella, A. Bestavros, Self-similarity in world wide web traffic: evidence and possible causes, IEEE/ACM Trans. Networking 5 (6) (1997) 835–846.

[6] M. Zukerman, T. D. Neame, R. G. Addie, Internet traffic modeling and future technology implications, in: Proc. IEEE INFOCOM 2003, Vol. 1, 2003, pp. 587–596.

[7] I. Norros, On the use of fractional Brownian motion in the theory of connectionless networks, IEEE J. Sel. Areas Commun. 13 (6) (1995) 953–962.

[8] R. S. W. Willinger, M.S. Taqqu, D. Wilson, Self-similarity through high-variability: statistical analysis of Ethernet LAN traffic at the source level, IEEE/ACM Trans. Networking 5 (1) (1997) 71–86.

[9] M. S. Taqqu, W. Willinger, R. Sherman, Proof of a fundamental result in self-similar traffic modeling, SIGCOMM Comput. Commun. Rev. 27 (2) (1997) 5–23.

[10] R. W. Dijkerman, R. R. Mazumdar, On the correlation structure of the wavelet coefficients of fractional Brownian motion, IEEE Trans. Inf. Theory 40 (5) (1994) 1609 –1612.

[11] N. G. Duffield, N. O'Connell, Large deviations and overflow probabilities for the general single-server queue, with applications, Mathematical Proceedings of the Cambridge Philosophical Society 118 (2) (1995) 363–374.

[12] I. Norros, A storage model with self-similar input, Queueing Syst. 16 (3-4) (1994) 387–396.

[13] I. Norros, Four approaches to the fractional Brownian storage, in: J. Lévy Véhel, E. Lutton, C. Tricot (Eds.), Fractals in Engineering, Spinger, 1997, pp. 154–169.

[14] O. Narayan, Exact asymptotic queue length distribution for fractional Brownian traffic, Adv. Perform. Anal. 1 (1) (1998) 39–63.

[15] L. Massoulié, A. Simonian, Large buffer asymptotics for the queue with fractional Brownian input, J. Appl. Probab. 36 (3) (1999) 894–906.

[16] J. Hüsler, V. Piterbarg, Extremes of a certain class of Gaussian processes, Stochastic Processes and their Applications 83 (2) (1999) 257 – 271.

[17] I. Norros, E. Valkeila, J. Virtamo, An elementary approach to a Girsanov formula and other analytical results on fractional Brownian motions, Bernoulli 5 (4) (1999) 571–587.

[18] Y. Jin, S. Bali, T. Duncan, V. Frost, Predicting properties of congestion events for a queueing system with fBm traffic, IEEE/ACM Trans. Networking 15 (5) (2007) 1098 –1108.

[19] A. Rizk, M. Fidler, Sample path bounds for long memory fbm traffic, in: Proc. IEEE INFOCOM 2010, 2010, pp. 1–5.

[20] J. Chen, R. Addie, M. Zukerman, Link dimensioning for fractional brownian input, in: Proc. IEEE HPSR 2011, 2011, pp. 151 –157.

[21] E. Reich, On the integrodifferential equation of takács I, Annals of Mathematical Statistics 29 (1958) 563–570.

[22] J. M. Harrison, Brownian motion and stochastic flow systems, John Wiley and Sons, New York, 1985.

[23] A. B. Dieker, B. Yakir, On asymptotic constants in the theory of gaussian processes (2012). arXiv:1206.5840 [math.PR].

[24] M. A. Lifshits, Gaussian Random Functions, Kluwer Academic Publishers, 1995.

[25] B. B. Mandelbrot, J. W. V. Ness, Fractional brownian motions, fractional noises and applications, Society for Industrial and Applied Mathematics 10 (4) (1968) 422–437.

[26] J. R. M. Hosking, Modeling persistence in hydrological time series using fractional differencing, Water resources research 20 (12) (1984) 1898–1908.

[27] C. R. Dietrich, G. N. Newsam, Fast and exact simulation of stationary Gaussian processes through circulant embedding of the covariance matrix, SIAM Journal Sci. Comput. 18 (4) (1997) 1088–1107.

[28] V. Paxson, Fast, approximate synthesis of fractional gaussian noise for generating self-similar network traffic, ACM Sigcomm Comput. Commun. Rev. 27 (5) (1997) 5–18.

[29] W. Whitt, Stochastic-process limits, Springer-Verlag, New York, USA, 2002, pp. 124–130.

[30] A. B. Dieker, Fractional Brownian motion.
URL http://www2.isye.gatech.edu/ adieker3/fbm/index.html

[31] R. G. Addie, T. D. Neame, M. Zukerman, Performance analysis of a Poisson-Pareto queue over the full range of system parameters, Computer Networks 53 (7) (2009) 1099–1113.

[32] R. G. Addie, T. D. Neame, M. Zukerman, Performance evaluation of a queue fed by a Poisson Pareto burst process, Computer Networks 40 (3) (2002) 377–397.

[33] R. G. Addie, On weak convergence of long-range-dependent traffic processes, Journal of Statistical Planning and Inference 80 (1-2) (1999) 155–171.

[34] H. Heffes, D. Lucantoni, A Markov modulated characterization of packetized voice and data traffic and related statistical multiplexer performance, IEEE J. Sel. Area. Comm. 4 (6) (1986) 856–868.

[35] D. Heyman, D. Lucantoni, Modeling multiple IP traffic streams with rate limits, IEEE/ACM Trans. Networking 11 (6) (2003) 948–958.

[36] The CAIDA UCSD Anonymized Internet Traces 2012 - [20120216].
URL http://www.caida.org/data/passive/passive_2012_dataset.xml

[37] M. W. Garrett, W. Willinger, Analysis, modeling and generation of self-similar VBR video traffic, ACM Sigcomm Comput. Commun. Rev. 24 (4) (1994) 269–280.

[38] P. Abry, D. Veitch, Wavelet analysis of long-range-dependent traffic, IEEE Trans. Inf. Theory 44 (1) (1998) 2–15.

[39] D. Veitch, P. Abry, A wavelet-based joint estimator of the parameters of long-range dependence, IEEE Trans. Inf. Theory 45 (3) (1999) 878–897.

[40] S. Bali, V. Frost, An algorithm for fitting MMPP to IP traffic traces.
URL http://www.ittc.ku.edu/ frost/NSF-QoS-Project/index3.htm